Practical Data Science
Assignment 2 Report

Dimitrios Koutsianos
f3352212

1 Introduction

In this short report we will present some of our findings for the 2nd Assignment of Practical Data Science. We will not present our methods here, only some figures and compare results in a short way as we have already discussed them in extent to the accompanying notebook.

2 Part 3

Here we compared the efficiency of three ML algorithms, namely SVM, K-NN and HistBoosting, on the classification task of our 2 datasets based on character, century and TM.

2.1 Character

The metric to calculate the efficiency of these models here is the *macro avg* since the characters classes are heavily unbalanced (as seen on Figures 1 and 2). The three models scored 59%, 54% and 48% respectively, which are not great percentages. Hence these models shouldn't be completely trusted. Although, if we tweak the parameters of HB, we should expect better scores.

2.2 Century

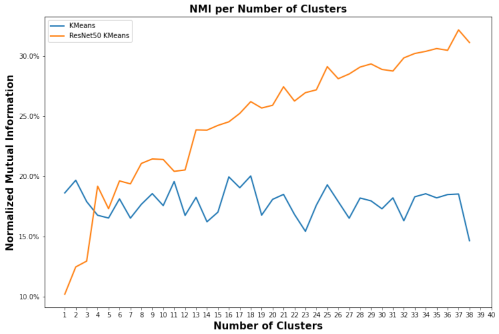
Here the *accuracy* metric is chosen since the 3 classes are mostly balanced (as seen on Figure 3). The three models scored 81%, 60% and 85% respectively. Hence, HB is the clear winner here, and a quite reliable one since 85% is actually a great score, which we expect to get even better with some tweaking.

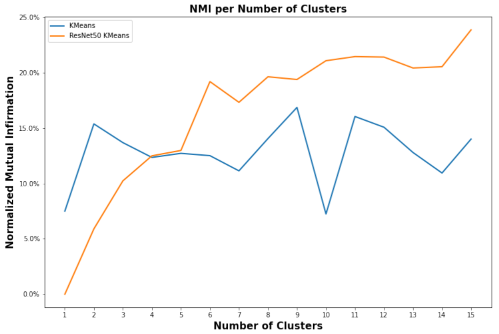
2.3 TM

Here the models were unimpressive scoring just 25%, 17% and 19% respectively on the *macro avg*, which we used because of the extremely unbalanced classes (see Figure 4).Thus these models are useless and completely untrustworthy.

3 Part 4

Here we performed clustering on the images of the 2nd dataset based on the character they represent and their century of origin. The metric we used to choose the optimal number of clusters was the Normalized Mutual Information Score and we used two different models on each clustering task, the usual KMeans and KMeans where we first used the pre-trained CNN *ResNet50* for feature extraction. The results are in the following two figures:





We clearly see that in both cases the *ResNet50* on is the better choice. The optimum k for each task is 35 and 15 respectively. But the problem of the low NMI remains, which can be attributed to the low sample per feature ratio, as we have close to 1000 samples for 784 features.

4 Part 5

For Part 5 we performed *agglomerative clustering* on each character class using the *Davies-Bouldin Criterion* with a range of 3 to 9 clusters for each character. The clustering was deemed successful for most characters, since it managed to find different features in each character and create a cluster of similarly written instances of each letter.

5 Appendix

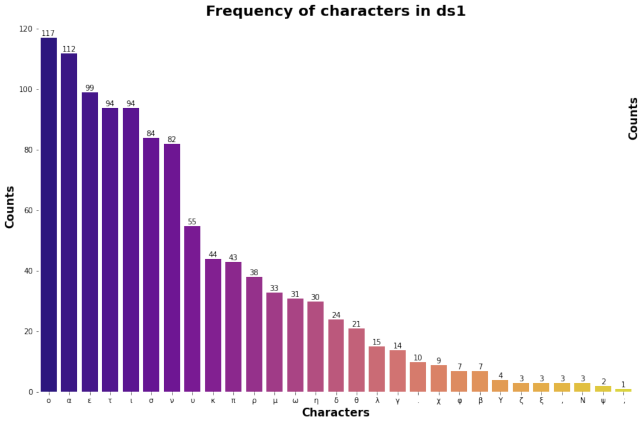


Figure 1: Character Frequency in *ds1*

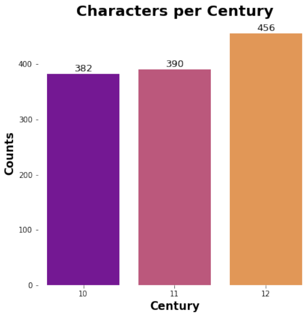


Figure 3: Century Frequency in *ds1*

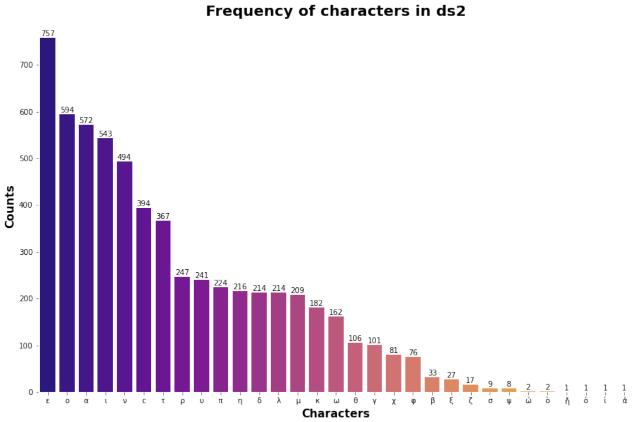


Figure 2: Character Frequency in *ds2*

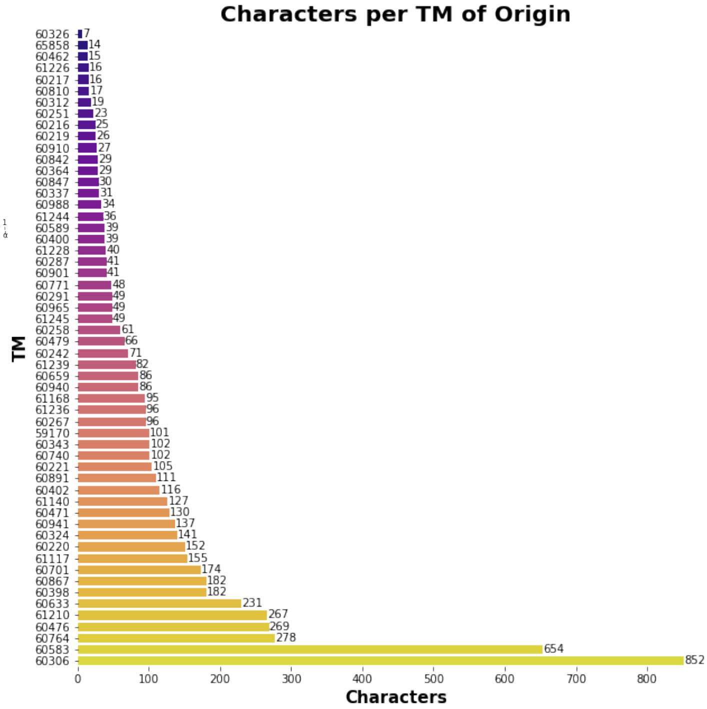


Figure 4: TM Frequency in *ds2*